DOCUMENT RESUME

ED 269 473 TM 860 330

AUTHOR Ridley, Dennis R.
TITLE An Analysis of the Uniform Student Evaluation Survey (USES).
PUB DATE 16 May 86
NOTE 30p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS College Faculty; Correlation; Curriculum Evaluation; Feedback; Higher Education; Instructional Improvement; *Item Analysis; Multiple Regression Analysis; Rating Scales; *School Surveys; *Student Evaluation of Teacher Performance; Teacher Effectiveness; *Test Reliability; Test Results; *Test Validity
IDENTIFIERS *Christopher Newport College VA; *Uniform Student Evaluation Survey

ABSTRACT

This report presents an analysis of a student evaluation survey, the Uniform Student Evaluation Survey (USES), used to provide a systematic, campus-wide basis for meaningful feedback on the instructional effectiveness of college teaching at Christopher Newport College (Virginia). This study addressed three issues as to the reliability, validity, and consistency of the survey measures. Secondarily, this report provided a device for guiding interested, but statistically non-sophisticated persons (primarily faculty), through the analyses of the three issues. A rationale for the current USES survey was presented and considerations for revision were offered. This report is limited to the analysis of 13 items (included in appendix) whereby a student evaluates an instructor's effectiveness. Principal components analysis of previous USES results revealed that "perceived classroom effectiveness" is consistently what is measured by the survey, with "overall instructor performance" the best single item measure. Items which shared the most variance with overall classroom effectiveness were similar from one department to another. USES results (1) show a high degree of internal consistency; (2) suggest that scores are fairly stable over time, with stability dropping sharply after two years; and (3) are generally related to more or less independent assessments of excellence as a classroom teacher. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED269473

An Analysis of the Uniform Student Evaluation Survey (USES)

Dennis R. Ridley
Assistant to the Vice President for Academic Affairs
Christopher Newport College
50 Shoe Lane
Newport News, Virginia 23606-2988

May 16, 1986

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. R. Ridley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 860 330

TABLE OF CONTENTS

	Page
PREFACE	i
EXECUTIVE SUMMARY	ii
MAIN REPORT	
Introduction	1
What does USES (PART A) measure?	2
Is USES (PART A) reliable?	14
How valid is USES (PART A)?	16
Summary of Major Conclusions	23
APPENDIX	
Uniform Student Evaluation Survey (PART A)	25

PREFACE

This report presents an analysis of the student evaluation survey which is currently in use at Christopher Newport College in Newport News, Virginia. Since 1981, this instrument, the Uniform Student Evaluation Survey (USES), has been used to provide a systematic, campus-wide basis for meaningful feedback on instructional effectiveness.

For the past two years, as the result of combined faculty-administration review and study of the USES, a number of findings have emerged. Although these results apply particularly to this survey, the approach taken as well as the results obtained are deemed to have much wider applicability. This study addressed several broad issues that similar reviews of any instrument, which are designed to provide information on instructional effectiveness of college teaching, also need to address. These include, primarily, the following questions: (1) What does the survey measure? (2) Is the survey reliable? (3) Is the survey valid?

A secondary purpose for this report was to provide a clear, systematic device for guiding interested, but statistically non-sophisticated persons (primarily, faculty), through the analyses of these three issues. The goal was to provide a rationale for the current USES survey, and guide considerations for its revision. Again, this report as a means to this goal, while specifically focussed on Christopher Newport College, hopefully will find broader applicability in serving as a model for similar applied research and educational efforts elsewhere.

This report is limited to the analysis of thirteen evaluative items which are the heart of the USES instrument. These thirteen items are listed as an appendix to the main report. Excluded from consideration in this analysis are a number of demographic questions which form the second part of the USES.

EXECUTIVE SUMMARY

What does the USES measure?

In addressing the question of what the USES measures, a method was selected to reduce the complexity found in the pairwise comparison among items on the survey. The method selected was principal components analysis (PCA).

The first analysis was on data derived from the 13 evaluative items of Part A of the USES. In all, these data were from 10 semesters, totalling 3,924 instructional performances, or courses/sections evaluated. The raw data matrix consisted of 51,012 separate pieces of information, i.e. average scores on 13 items, times 3,924 evaluations. Two distinct factors were found. The first, accounting for about 65.1 percent of the variation in student ratings, was correlated with questions which largely pertain to the impression or impact of the instruction within the classroom. The second factor pertained to outside-of-class indicators of instructor performance.

To a large degree, these results were found to be stable from one semester to the next and across each of the 18 departments of the college. Therefore, the structure of the survey is not strongly affected by local idiosyncrasies or temporally specific factors. The general pattern of results is less stable from one individual to the next showing that these patterns are most consistently shown when averaging across individuals.

At the item level, it was discovered that the item which most consistently had the highest correlation with the first, primary component was item number 13. This item asks for the student's summative rating of the instructor's overall performance as an instructor, using a five-point scale: "unacceptable," "weak," "satisfactory," "strong" and "outstanding." Among other items which correlated highly with the first factor were item numbers 11 (motivation/stimulation), 9 (clear presentations) and 2 (well-prepared). Multiple regression analysis revealed that generally the same items, particularly number 11, best predicted the average ratings of overall performance.

How reliable is the USES?

An estimate of the internal consistency reliability of the USES survey was derived by means of Cronbach's alpha coefficient. Since the first and primary principle component accounted for the majority of the variation among ratings, and most items correlated above .8 with this factor, it was decided that serious distortion would not occur by including all of the 13 evaluative items in this analysis.

When this analysis was carried out, an extremely high alpha coefficient of .95 was found. At the same time it was found that deleting any one of the 13 items from the total scale reduced alpha by less than .01. Thus, we may reasonably conclude that Part I of the USES is as internally consistent as it needs to be for the purposes for which it is used.

A related question is how stable the survey ratings are over time. To address this question, an analysis of means for item number 13 (overall performance) was carried out. This was intended to provide an efficiently calculated, and conservative, estimate of stability. The data were based on 90 instructors for whom annual summaries of USES scores were available for the previous four years.

Three pairwise comparisons between score distributions, separated by one, two, and three years, were carried out by means of Pearson correlations. The resulting correlations were .84, .77, and .59, respectively. These results suggested that scores are fairly stable over time, with stability dropping sharply after about two years.

How valid is the USES?

The crucial question addressed here was distinguished from that of face validity of the items. It was whether USES data are "valid enough" for the various uses to which they are, or might be, put. The two primary purposes are as input for evaluation of faculty and as a diagnostic tool for the improvement of instruction. Most of the attention was directed toward the issue of validity in terms of the first purpose since this is the purpose that is foremost in the minds of faculty evaluees.

The relevant data presented in this report concern how USES results compare with professional peer assessment of teaching performance, or other reflections upon teaching which are independent of (i.e. not influenced by) USES results. These data were the mean overall performance ratings of eleven members of the faculty who received any one of several awards for distinguished teaching. These awards were made independently of USES scores; however, the average overall USES scores were about one standard deviation above the all-college average. There was also some evidence that low average scores on USES were related to the probability of termination.

Although some evidence of validity was found, it was also stressed that this evidence was based on aggregate results. The potential for errors in any particular individual's case was significant. Therefore it was recommended that while USES averages might have a corroborative function when considered along with other indicators of teaching effectiveness, they should under no circumstances be the sole arbiter in any key personnel decision regarding faculty.

INTRODUCTION

This report represents an attempt to answer a few fundamental questions regarding the USES (Uniform Student Evaluation Survey) instrument. This instrument consists of three parts: 1) thirteen questions directed toward evaluation by students of their instructor in terms of his or her effectiveness; 2) seven questions, ostensibly permitting one to investigate the size and direction of relationships (if any) between the students' evaluations of instructors and a number of student characteristics such as age, sex, expected grade in the course, etc; and 3) five questions of students' judgments relative to course difficulty, appeal, value to the student, the students' motivation to succeed in the course, and the effectiveness of the survey itself.

This report has only a limited objective, but an extremely crucial one: that is to assess the first part of the survey (i.e., the thirteen questions on perceived instructor effectiveness) in terms of a few technical questions that might also be asked, in one form or another, of virtually any test or survey.

The most general of these questions are essentially three:

1. What do the questions or items comprising the survey actually (as opposed to intentionally) measure?

2. Is the survey a reliable instrument, or consistent and accurate in measuring whatever it does measure?

3. Is the survey a valid instrument in that what it actually measures is in accordance with the purposes for which it was designed?

The reader will recognize that these three questions are closely and logically related to each other. For the instrument to have any value whatsoever, we must first be able to discover empirically what is being measured, whether that be simple or complex and whether or not it measures what was originally intended. Once we have ascertained that to our satisfaction, we must ask whether the measure is consistent and accurate in that measurement, or in those measurements (if the instrument is complex). This notion of reliability is fundamental, and is also complex. First, note that reliability is a matter of degree, and an index of reliability reflects this by varying from zero to one. If the reliability approaches zero, it should be manifest that since the instrument practically lacks all consistency and accuracy, nothing is being measured at all. If the reliability approaches one, we have a highly reliable instrument but not necessarily a valid one, in the same sense that a marksman may shoot consistently at a target, and consistently hit above or below the target when the sights are off. Reliability also is complex in the sense that we may be interested in at least two different senses of accuracy: 1) How consistent is one part of the measure or scale with another part? 2) How stable is the instrument in giving you the same results at different times? We shall examine both types of reliability in this report.

Finally, the validity of the instrument is perhaps the most crucial question of all, and yet the answer to the question, "How valid is the instrument?" depends on the other questions in definite ways. The instrument must first measure something which can be discovered empirically and described. But what it measures may or may not bear any resemblance to the intended purpose or purposes of the instrument. This degree of correspondence to intended purposes is what we mean by validity. Obviously, it is also a matter of degree, and any

index of validity or data bearing on validity should help one to judge whether the validity achieved is sufficient for the uses desired for the instrument. Validity also depends upon reliability, as we have already observed with the target analogy. But there is also hidden complexity in this notion of validity just as there is subtlety in the concept of reliability. If validity is relative to the intended purpose or purposes of the instrument, there may need to be more than one assessment of validity since there may be more than one intended purpose. Or, what really amounts to the same thing, the intended purpose may, upon careful analysis, turn out to be more complex than was originally supposed. The assessment of validity should reflect this fact.

WHAT DOES USES (PART A) MEASURE?

Quite independently of the intended purposes of the designers of USES (Part I) to assess students' perceptions of instructors' effectiveness, USES could be complex or simple in terms of what we discover is measured by the various items. We can address this issue by investigating the relationships among the items to see what response patterns they evoke from a variety of students when evaluating different professors or different courses. For example, let us suppose most students consistently give Instructor A a relatively high rating on the item, "The instructor was intellectually motivating and stimulated learning" (item no. 11). Let us further suppose that most students give Instructor B a relatively lower rating and give Instructor C a rating which is intermediate between A and B on the same question. These patterns do not necessarily mean that these students give the same ratings, but only that overall, their respective ratings on that item show that they tend to rank A, B and C in the same relative positions. Now let us suppose that when we examine another item (e.g., no. 9: "The instructor's presentation of the course material was clear and understandable.") we were to find almost exactly the same patterns of ratings by students for the same instructors, A, B and C. We would probably be justified in making the generalization that instructors whose presentations are intellectually stimulating and motivating are also instructors whose presentation of course material is clear and understandable, and instructors who are lacking in one tend to be lacking in both. (Can we really imagine an obscure and incomprehensible presentation of course material being intellectually stimulating?) Similarly, we can continue to examine the relationships among items in terms of rating patterns, and we would begin to find other plausible generalizations. We would probably conclude that, although there are 13 separate questions, there is actually a great amount of overlap or redundancy among the various items in terms of what is measured.

This is actually the sort of thing we have attempted to do. But the problem with examining, item by item, the pairwise comparisons among them in the patterns of ratings to which they give rise, should be obvious. There are far too many of these relationships to examine and keep in one's mind. The actual number of pairwise comparisons in n items is given by:

$\frac{n!}{2\,(n-2)\,!}$ which is, for 13 items, 78. This number of comparisons is quite unmanageable.

For this reason, techniques called principal components analysis and factor analysis have been devised. Without going into technical matters to define these techniques, suffice it to say the following. These are mathematical (sta-

tistical) techniques by which to extract meaningful patterns among the various items which make up a test or survey. These patterns are interpreted empirically, that is, by reference to the item or items which correlate most highly with each factor, or underlying variable extracted form the data. Such factors are not necessarily typified by any one item on the test, but by looking at the several which are most representative of the factor or factors, we can offer plausible explanations of what the test appears to measure. There may be, of course, more than one factor. This speaks to the issue of the simplicity versus complexity of the instrument.

Principal Components of USES

A principal component analysis was conducted over ten past semesters (not including the present semester) during which the USES was administered. We also selected two semesters to examine whether the patterns observed for all past semesters are consistently replicated from one semester to the next. The same analysis is also repeated, separately, in all 19 departments at Christopher Newport College. The purpose of this was to assess whether there are local exceptions to the college-wide results. Finally, analyses for three individual faculty members, one from each school of the College, were carried out separately. The purpose of these analyses was to assess individual variations from the broader patterns for the instructors' departments and for the College as a whole.

In all of these analyses, the unit of analysis is the course or section rather than the instructor. Each course (or section taught by a single instructor) can be thought of as a performance by that instructor. On this dramaturgical analogy, the student raters using the USES would be the reviewers of the performance. Just as actors give many separate performances and are reviewed as often as they perform, so instructors generate a variable number of occasions for evaluation depending on how many courses they have taught during the ten semesters which are included in this analysis.

In all 10 semesters there were 3,924 individual performances, or courses taught. For each course, a variable number of students rated the instructor on all 13 items of Part I, and an average score on each of these items was computed. These are the same results as are generated for faculty members' course evaluations, except that here, the aggregated data were to be analyzed. A total of 51,012 separate pieces of data (each one an average rating on one of 13 items for one of 3,924 distinct courses) were analyzed.

First, the results over ten semesters are presented. These results (Table 1) show two distinct factors. As mentioned above, each factor can be interpreted by reference to the items that correlate most highly with the factor. To aid this interpretation, Table 1 is organized with the questions which most typify the factor listed first, the remainder of the items following in descending order. The number given is the correlation between the item, listed by number and brief description, and the factor. Each factor has been given a label to suggest the general meaning implied by the items that correlate with it.

Table 1. Factor Analysis of USES (Part I: Ten Semester - Results)

Item No.	Item-Factor Correlation	Brief Item Description
Factor 1. Classroom Effectiveness		
11	.87	Motivation Stimulation
10	.87	Interest/enthusiasm
8	.85	Expertise
13	.84	Overall instructor performance
9	.82	Clear presentations
2	.79	Well-prepared
Factor 2: Grading practices		
5	.82	Prompt grading
4	.73	Fair grading
3	.72	Clear grading practices

Factor 1 is a broad factor which accounts for about 65.1 percent of the variation in student ratings. The items that correlate most highly with the factor largely pertain to the impression or impact created by the instructor within the classroom. The specific items that correlated highly with Factor 1 can be deciphered from their brief descriptions in Table 1. (In the interest of total clarity, Part 1 has been reproduced verbatim as Appendix 1.) They suggest that instructors who rate high on this factor are stimulating, enthusiastic, knowledgeable, clear and well-prepared.

In contrast, Factor 2 pertains to outside-of-classroom indicators of instructor performance, in particular grading practices. The specific items can be read in their entirety in Appendix 1.

These two factors, while distinct, are not necessarily unrelated for particular instructors. Indeed, they tend to have a modest positive relationship overall. That is, the items that most typify Factor 1 are uniformly positively (but modestly) correlated with those that are most representative of Factor 2. However, it is not too difficult to imagine instructors being highly rated on Factor 1 items whose grading practices come in for some criticism, or instructors who are perceived to be ineffective in the classroom being perceived as scrupulously clear, fair and prompt in their grading. It is much less imaginable that instructors would be perceived as high on some items that measure Factor 1 and low on others. This coherence among perceptions that are reflected in the items comprising a factor is the reality behind the identification of a factor as such.

Turning to the results for one semester, we find a similar pattern of results. The first semester studied was the Fall of 1981. Table 2 shows the results of this analysis.

Table 2. Principal Components Analysis of USES (Part 1): One Semester Results - Fall, 1981

Factor 1. Classroom Effectiveness

Item No.	Item-Factor Correlation	Brief Item Description
11	.89	Simulation/motivation
13	.88	Overall instructor performance
10	.86	Interest/enthusiasm
9	.85	Clear presentations
8	.85	Expertise
2	.70	Well-prepared

Factor 2. Grading Practices/Administrative Effectiveness

5	.84	Prompt grading
3	.67	Clear grading practices
4	.62	Fair grading

This analysis of 474 courses again turns up two factors. The first, accounting for 61.4 percent of the variation in average responses, has essentially the same six items most highly correlated with it. The minor differences between this semester and the overall results are confined to differences in the relative ordering of these six items. The overall impression created by this factor is that it relates directly to classroom performance. The second factor, accounting for 8.3 percent of the variation in ratings, is almost identical to the second factor for the overall result. Again it refers to grading practices.

The second individual semester analyzed again turned up essentially identical results.

Table 3. Principal Components Anaysis of USES (Part 1): One Semester Results - Fall, 1984

Factor 1. Classroom Effectiveness

Item No.	Item-Factor Correlation	Brief Item Description
13	.93	Overall instructor performance
11	.91	Stimulation/motivation
2	.90	Well-prepared
9	.90	Clear presentations
8	.89	Expertise
10	.88	Interest/enthusiasm

Factor 2. Grading Practices

5	.63	Prompt grading

[No other items were well correlated with Factor 2.]

This analysis of 523 courses gives additional evidence for the factors. The first accounts for 68.3 percent of the variation in average responses, and the second explains just 6.6 percent of this variation. The first factor has the same flavor as shown in the above analyses. It measures perceived excellence in performance within the classroom. Again, the only variation shown is in the relative ordering of items, but this is of little significance. We may conclude that the results do not vary substantially from one semester to another.

Exactly the same principal components analysis was carried out for each of the 18 departments for all 10 semesters combined. To summarize the major results, Table 4 below compares six items in terms of their degree or correlation with Factor 1, Classroom Effectiveness. Instead of correlations, the ranks of these correlations within each department are given. For example, in the Department of English, the items in order of strength of correlation, from 1 to 6, are items 13 (overall effectiveness), 11 (motivation and stimulation), 9 (clear and understandable), 2 (well-prepared), 8 (expertise), and 10 (interest and enthusiasm). This order of items for the Department of English turns out to be "typical" of the departments as a whole in the sense that averaging the within-department ranks over all departments results in the same order. Also note, from Tables 1-3, that the same six items correlate most highly with first factor when the principal components analysis is done over all departments, although the order is somewhat different.

Table 4. Items Most Strongly Correlated With Factor I (Classroom Effectiveness) Ranked Within Departments

	USES ITEM NOS.					
Department	13	11	9	2	8	10
B.C.E.S.	2	1	3	4	11	6
C.S.	1	2	4	3	6	5
ENG.	1	2	3	4	5	6
MATH	1	3	2	5	6	4
M.C.L.L.	1	3	2	5	4	8
PHIL.	1	2	4	9	8	3
PHYS.	2	1	4	3	5	6
ACC./FIN.	1	4	3	2	8	5
ECON.	1	2	4	5	7	3
M./M.	1	3	2	4	5	7
EDUC.	1	2	3	6	7	4
A./C.	1	2	9	4	3	5
HIST.	1	4	3	8	2	6
L.S./P.E.	1	2	3	5	8	7
M.S.	4	1	5	10	2	8
P.S./G.A.	1	4	2	3	5	7
PSYCH.	2	4	1	3	6	8
SOC./S.W.	1	2	4	6	3	7
Mean Rank:	1.33	2.44	3.39	4.94	5.61	5.83

What is the most striking from Table 4 is the general consistency of the results over the departments. For example, among the three items with the highest average within-department ranks (13, 11 and 9) there was only one instance of a rank greater than 5. Although there was more variation in ranks

among the lowest ranked items, there was also a degree of consistency. For example, in no department was any of the three lowest ranked items correlated most strongly (rank = 1) with Factor 1.

The first principal component found within departments is quite similar to that found over all departments. These analyses show that the best single measure of this factor is item no. 13, overall effectiveness, which in 14 out of 18 departments had the highest rank in terms of correlation with the factor.

The primacy of Factor 1 is again revealed in the percent of variation which is accounted for in each of the departments. Factor 1 explains between 54 and 76 percent of the variation with the majority being over 65 percent. This corresponds fairly well with the results over all departments combined. The second factor in each department explains between 6 and 13 percent of the total variation. Although this result is also consistent with the results for all departments combined, a much less consistent view of the significance of the second factor emerges. The most that can be said for certain is that in each department the second factor to emerge is correlated most strongly with a non-classroom-related indicator of effectiveness, such as prompt grading, fair grading, punctuality, and availability to the student. However, beyond that, there is no clearly consistent significance of a second factor.

Principle components analyses were also done on the ratings on Part 1 items for three individual instructors, one from each school of the College. These instructors were chosen at random from among those who had evaluations for at least 20 courses. These analyses, summarized in Table 5, were done to assess the degree of consistency of factors across persons.

Table 5. Principal Components Analyses for Three Instructors

Instructor A		Instructor B		Instructor C	
Item	Item-Factor Correlation	Item	Item-Factor Correlation	Item	Item-Factor Correlation
9	.92	9	.88	10	.91
11	.91	11	.87	9	.90
3	.90	10	.86	1	.88
13	.90	6	.86	2	.86
2	.87	7	.84	11	.86
10	.86	13	.83	8	.83

In contrast with the results presented earlier, we find somewhat less consistency in the results across individual instructors' data. It is at least clear that there is one major factor in each instructor's ratings. This factor accounts for over 50 percent of the total variation in each instructor's ratings over all items. It is also clear that what is being measured is quite strongly correlated with effectiveness of the instructor within the classroom (items 9, 11, etc.) However, there is apparently also a good deal of individual variation shown by the fact that other items, not directly related to classroom performance (e.g., items 1, 3, 6, and 7) may be as highly correlated with Factor 1 as the items which are directly related to classroom performance. These individual variations tend to disappear when we take in the larger, composite picture. What we have identified as Factor 2 is also less clear on the individual professor level.

To review the overall picture, a remarkably consistent pattern emerges in which there is one major factor which accounts for approximately 70 percent of the variation. This factor is best interpreted as overall classroom effectiveness, as suggested by three of the best single item measures of this factor which are item no. 11 (motivation/stimulation), item no. 13 (overall performance), and item no. 9 (interest/enthusiasm).

There is also a much less important factor, in terms of variation accounted for, which explains roughly 7 percent of the variation. The meaning of it is also less consistent and clear. What is at least clear is that the second factor tends to correlate with items having to do with grading practices, punctuality, availability, and the like, which do not bear directly on classroom performance or its effectiveness as perceived by students.

These patterns are consistent and stable from one semester to the next. They also do not vary greatly from department to department, except that the weaker second factor fluctuates in terms of the outside-of-classroom indicators of performance that it correlates with. Finally, the general pattern of results is less stable from one individual to the next showing that the patterns we have noted are most consistent when we average across individuals.

Analysis at the Item Level

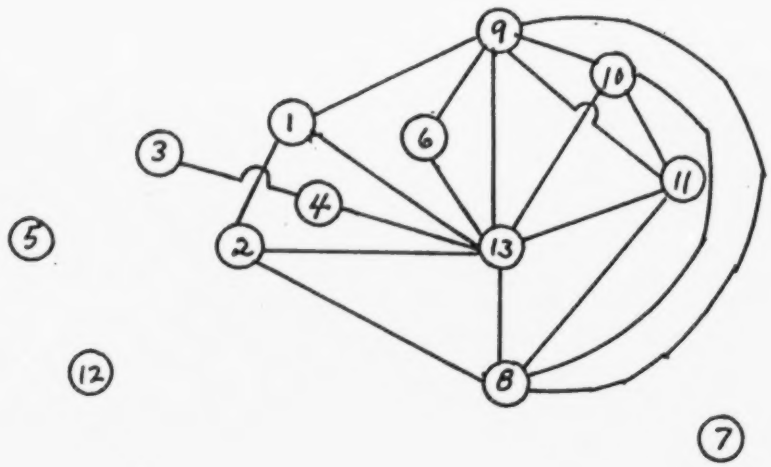
Item no. 13, which indicates how students perceive instructors' overall performance, is arguably the best single indicator of the first factor, measuring student perceptions of classroom effectiveness. In the separate factor analyses by departments, item 13 had the highest correlation with the primary factor 14 out of 18 times. In the other four departments, the best correlating items were item 11 (motivation/stimulation-3 times) and item 10 (interest/enthusiasm-1 time). These frequencies strongly suggest that it is item 13, and secondarily item 11 (motivation/stimulation) and item 9 (interest/enthusiasm) that best define the primary factor accounting for the variation in student ratings of instruction (see Table 4).

The separate correlations among test items tell essentially the same story, although less clearly. All of the 13 items comprising Part I of the survey are positively correlated with each other. However, the three items with the highest correlations are items 11 and 13 (.90), items 9 and 13 (.85), items 10 (clear and understandable presentations) and 11 (.85), and items 9 and 11 (.83). These results over all ten semesters are essentially repeated, with very minor differences, in individual semesters.

A graphic presentation of correlations among the USES items can reveal a network of relationships like that suggested by the principal components analyses. In the following figure, numbers inside circles correspond to item numbers on the survey. Lines between these circles indicate a correlation greater than .7. With a correlation of at least .7, we know that from a knowledge of students' ratings on one of a pair of variables, the ratings on the second member of the pair can be predicted with at least fair accuracy. For example, as mentioned earlier if we know how students rank instructors in terms of clarity of presentations, we can be sure that they will rank instructors in almost the same way in terms of how stimulating they are. Therefore, these lines between items can suggest the commonality among items in almost the same way that principal components or factor analysis does.

In Figure 1, the reader will note the clear centrality of item no. 13, as also was suggested by the principal components analysis. The reader will also find evidence of a strong interlocking network among items 9, 10, 11 and 13, which are among the best indicators of the major factor. It is also apparent that there are very few outlying items, i.e. items which do not correlate at least .7 with one or more other items. It is instructive that these outlying items tend to be the ones which appear on the second factor, i.e. item 5 (prompt grading), item 12 (punctuality and time use), and item 7 (availability and office hours).

Figure 1. Empirical Relationships Among USES items. (Numbers represent USES item numbers; lines represent Correlations $\geq .7$.)



Since item no. 13 appears to be quite central to students' perceptions of instructors' classroom performance as measured by the survey, and since this pattern hardly varies from one department to another, we may simply look at item 13 by itself to get a further quick look at how the various questions are related to perceived overall performance. Although some of the following is already implied by the above discussion, for those who are not accustomed to dealing with correlations, the following should make these findings somewhat more concrete. For example, if there is a low correlation between two items, such as item 5 (prompt grading) and item 13 (overall performance), we can surmise that it is quite possible for an instructor's performance to have a high overall rating even when he or she is given a low score for grading promptness (or vice versa). We also surmise that a high correlation between another item and item 13 means that this pattern of discrepant responses is rather more unlikely (e.g., we can hardly imagine the same instructor being perceived as highly stimulating but low in overall performance, or vice versa). In essence, students are much more forgiving of certain practices or traits than others in terms of their impact on perceived overall performance. While this much is clear, these points can be illustrated more concretely by examining the actual pairwise responses of students, taking item 13 with each of the other items of Part I in turn.

The following tables (6, 7 and 8) give the percentages of the total number of respondents (over 10 semesters and nearly 66,000 ratings of instructors' performances) who answered item 13 with each of 5 possible responses and other

items (nos. 11, 9 and 2) with each of the possible ratings. Tables such as these are sometimes called "contingency tables" because they show how respondents answered one item contingent upon how they answered another. There are 20 possible ways in which pairs of responses may be recorded, i.e. 5 rating points X 4 rating points. In these tables, percentages rather than frequencies are given in order to afford an easier comparison with subsequent results for individual departments. Three items only are paired with item 13. These are the three which, along with question 13, were the strongest indicators of the first factor (see Table 4). Since the results are so similar they are presented together.

Table 6. Percentages of Respondents Giving Each Possible Pair of Answers to Questions 11 and 13

Question 13 (Overall Instructor Performance)

Question 11 (Stimulation/Motivation)

	Unacceptable	Weak	Satisfactory	Strong	Outstanding	
Strongly Disagree	1.0	2.0	1.2	0.3	0.1	4.7
Mildly Disagree	0.2	1.8	5.1	2.3	0.4	9.8
Mildly Agree	0.1	0.7	6.7	17.0	5.1	29.5
Strong Agree	0.1	0.3	2.2	15.8	37.6	56.0
TOTAL:	1.4	4.8	15.2	35.4	43.2	100.0

Table 7. Percentages of Respondents Giving Each Possible Pair of Answers to Questions 9 and 13

Question 9 (Clear & Understandable)

	Unacceptable	Weak	Satisfactory	Strong	Outstanding	
Strongly Disagree	0.9	1.8	1.0	0.3	0.1	4.2
Mildly Disagree	0.2	1.8	4.6	2.3	0.4	9.3
Mildly Agree	0.1	0.9	6.9	14.9	5.0	27.8
Strongly Agree	0.1	0.3	2.8	18.0	37.6	58.7
TOTAL:	1.3	4.8	15.3	35.5	43.1	100.0

Table 8. Percentages of Respondents Giving Each Possible Pair of Answers to Questions 2 and 13

Question 2 (Well Prepared)	Unacceptable	Weak	Satisfactory	Strong	Outstanding	
Strongly Disagree	0.5	0.5	0.2	0.1	0.1	1.3
Mildly Disagree	0.4	1.2	1.3	0.4	0.1	3.5
Mildly Agree	0.4	2.0	6.7	7.4	1.7	18.1
Strongly Agree	0.2	1.1	7.1	27.5	41.3	77.2
TOTAL:	1.4	4.8	15.3	35.4	43.2	100.1

In the above three tables the total column and row percentages are only of secondary interest. However, the generally positive ratings on all four of these items are evident in the increasing total column percentages as one reads from left to right, and the corresponding increases in total row percentages, reading from top to bottom.

What is most relevant to discuss here is the relative symmetry of each of these tables. Granted that the majority give fairly high ratings, what these tables reveal is that lower ratings on items 11, 2 or 9 are accompanied by lower ratings on item 13. The extreme off-diagonal responses are quite rare. (Perfect symmetry is impossible since item 13 has one more rating point than the others have.)

Although other item pair comparisons are not reported, it should suffice to say that, by inspection and by several statistical tests, these three show the greatest degree of symmetry in paired responses of all combinations with item 13.

The above analysis provides information regarding how important each of several key items is to perceived overall instructor performance. However, this information only pertains when these items are considered singly and in isolation from other items. This involves an artificial exclusion of part of the available evidence. Fortunately, we have the means to go further and ask how the items, considered jointly and in combination, can predict perceived overall performance. The method of choice is termed "multiple regression analysis."

In a multiple regression analysis, the goal is to derive an equation that predicts with the least possible error, some measurable criterion. In this case the criterion is perceived overall instructor performance and the measure of it is responses on item 13. Multiple regression allows one to correct for redundant information in making the optimal prediction of the criterion.. For example, we already know that items 2, 9 and 11 are highly correlated with this criterion. However, we also know that they correlate quite well with each other. Therefore, to a certain degree they provide redundant information in that what is contributed by one item to the prediction is, to a large extent,

also contributed by the others, and vice versa. Multiple regression produces an equation in which the separate terms of the equation represent one item's capacity to predict a portion of the criterion which supplements the capacity of other items to predict other parts of the criterion. We know that the criterion, perceived overall instructor performance, is complex. We also know that the various items which correlate with it, though related to each other, are not identical and are imperfectly correlated. Thus, we expect and find an equation, through multiple regression, which is a linear combination of item variables. The equation takes the form:

$$V(13) = (a)(Vp) + b(V_q) + \ldots + K,$$

where V(13) is the predicted rating on variable (item) 13, a and b are numerical weights, Vp and Vq are scores on the first and second items to enter into the equation, and K is a constant.

The multiple regression analyses were done across all 10 semesters for all departments (N = 3900 +) and for each department separately. The purpose was to determine which items combine to make the most accurate prediction, and to see how much these patterns vary over the departments. For this analysis, all items in the survey were tested for possible entry into each equation. This provides a broad test of the relative importance of items in Part 1 and in other parts of the survey to the predictions.

Table 9 presents the regression equations for the entire college and for each academic department. The overwhelming majority of items in these questions are from Part 1. This shows the clear primacy of the evaluative items over the items bearing upon student characteristics and student judgments not directly related to instructor evaluation. Also of note is the fact that the first item to enter into the equations is almost invariably item 11. The second item to enter in is more likely to be item 2 or item 9 than any other item. The first two items invariably account for nearly all of the predictive value of the equation. For example, in the all-college prediction of item 13, a 2-variable solution (item 11 and 2) would have resulted in a multiple correlation of nearly .92. With a 4-item solution, the multiple correlation rises only a point to .93. This result is similar to that which is discovered for each of the departments considered individually. Therefore, differences between the equations beyond the first two items are more or less academic. The local variations from the all-college result appear to be minor also because even the second variable typically adds only a point or two to the multiple regression. Thus the results portray broad consistency over all departments.

Table 9. Regression Equations and Multiple Correlations for CNC and Separate Departments

UNIT	REGRESSION EQUATIONS	MULTIPLE CORRELATION
ALL-COLLEGE	V13 = (.56)(V11) + (.23)(V2) + (.19)(V4) + (.21)(V9) + (.15)(V1) - 1.53	0.93
ACCT	V13 = (.34)(V9) + (.48)(V11) + (.34)(V2) + (.18)(V3) + (.004)(V27)* - 1.63	0.94
ECON	V13 = (.55)(V11) + (.23)(V4) + (.15)(V2) + (.21)(V9) + (.27)(V1) - 1.77	0.97
MGMT	V13 = (.56)(V11) + (.32)(V9) + (.14)(V3) + (.28)(V2) + (.13) (V4) - 1.88	0.95
BCES	V13 = (.38)(V11) + (.40)(V9) + (.21)(V4) + (.20)(V8) + (.32)(V15)- 1.25	0.92
CPSC	V13 = (.39)(V11) + (.44)(V9) + (.25)(V2) + (.13)(V7) + (.14)(V1) - 1.52	0.97
ENGL	V13 = (.58)(V11) + (.21)(V3) + (.15)(V5) - (.55)(V18) + (.17)(V4) - 0.62	0.91
MATH	V13 = (.43)(V11) + (.33)(V9) + (.25)(V6) + (.26)(V2) + (.06)(V26)** - 1.33	0.93
MCLL	V13 = (.33)(V9) + (.40)(V11) + (.29)(V6) + (.39)(V8) - (.19)(V16) - 1.66	0.92
PHIL	V13 = (.30)(V11) + (.42)(V2)+ (.35)(V9) + (.18)(V23) - 1.20	0.92
PHYS	V13 = (.63)(V11) + (.19)(V9) + (.24)(V1) + (.24)(V6) - 1.31	0.93
A&C	V13 = (.66)(V11) + (.40)(V2) + (.28)(V4) - 1.58	0.88
EDUC	V13 = (.72)(V11) + (.40)(V9) + (.20)(V5) + (.07)(V13) + (.24)(V24) - 2.56	0.97
HIST	V13 = (.53)(V11) + (.47)(V2) + (.17)(V4) + (.20)(V6) - 1.67	0.93
LSPE	V13 = (.66)(V11) - (.25)(V20) + (.24)(V4)+ (.19)(V2) + (.11)(V25) - 0.94	0.93
POLS	V13 = (.32)(V11) + (.30)(V9) + (.24)(V23)+ (.32)(V2) + (.17)(V19) - 1.44	0.90
PSYC	V13 = (.70)(V11) + (.30)(V1) + (.26)(V4) + (.13)(V12) - 1.78	0.90
SOSW	V13 = (.39)(V11) + (.39)(V8) + (.26)(V9) + (.22)(V6) - 1.33	0.92

* This variable is the no. of respondents.

** This variable is the course level.

Interestingly, the same items which were found to be most important, when considered singly, are still the most important when variables are considered jointly. These are items 11, 9 and 2. One may conclude that instructors generally are judged to be doing an effective job if they are perceived as motivating and intellectually stimulating, if their presentations are clear and understandable, and if they are well-prepared. Beyond this, other items contribute relatively little to the overall impression that students have of instructors' effectiveness.

Multiple correlations, found in the extreme right hand column of Table 9,

average well over .90 for all departments. The multiple correlation for all departments combined is .93. These correlations indicate a very high level of predictability of the overall ratings of instructors' effectiveness. Although these results consider every variable having any significant contribution, which are included in the regression equations, the magnitude of these correlations is reduced only very slightly by limiting the number of predictive variables to the two or three which make the highest multiple correlations. These three strongest predictive variables invariably include item 11, include item 9 about half the time, and include item 2 about one-third of the time.

IS USES (PART A) RELIABLE?

Having examined in detail the question of what USES (Part I) measures, it is appropriate to ask how consistently and accurately it measures what we have already discovered is being measured. Note that some degree of reliability is implicit in what we have already found, since if reliability were close to zero, we could conclude that virtually nothing is being measured at all--that all the scores reflect is random noise. This is manifestly not the case. There is a meaningful and interpretable latent structure in the pattern of the scores, and this pattern, at least, is found quite consistently. The question remains, how consistent and accurate--i.e. how reliable--are these measurements? In particular, are they reliable enough to form one basis--although not the sole basis--for decisions relative to an individual's instructional contributions?

This question is complicated in that we have discovered that USES (Part I) itself is complex. We have found that there are at least two factors (or principal components). Therefore, the issue becomes two questions: How reliable is the measurement of Factor 1, and how reliable is the measurement of Factor 2?

Internal Consistency of USES (Part A)

Since Factor 1 is by far the most important factor, accounting for the majority of the variation in the previous analyses, an estimate of the reliability for this factor can be found by treating Part 1 as if it were all one scale. The question here can be stated as follows: To what extent is one part of the measure or scale (Part 1) consistent with another part? Since all items with the exception of item 5 (prompt grading) correlate above .6, and most correlate above .8 with Factor 1, the inclusion of all items will not result in a serious distortion in the estimate of internal consistency. Indeed, the estimate of reliability will be conservative since a "purer" scale should produce a higher estimate.

The statistic selected as the estimate of internal consistency is called Cronbach's "alpha coefficient." When one analyzes a large array of USES ratings, the total amount of variability within that array can be separated into different parts, or sources. In this particular analysis, there were 51,012 separate ratings (3924 performances x 13 variables in Part 1). The main sources of variability among ratings, in a statistical sense are three: 1) variability which is due to "systematic" differences among the 13 variables or items; 2) variability which is due to "systematic" differences among the 3,924 performances; and 3) "non-systematic" differences. The operative word here is "systematic." This refers to the fact that there are knowable differences in ratings, which are due either to performances or variables, and the data will reveal what these differences are. In the case of performances, we find that some courses receive

higher ratings (summed over all 13 variables), some receive lower ratings, and some receive ratings which are somewhere in between. (What we cannot separate in this analysis is how much of these differences can be attributed to instructors, and how much to courses.) In the case of the 13 variables, we already know that the variables are somewhat diverse in the attributes or traits that they tend to measure. How much of the total variability is due to each of these sources can be determined. The "non-systematic" variability is simply that part of the total variability that remains after the total has been partitioned and the systematic portions removed. It is due to an innumerable variety of contingencies that can affect a student's ratings apart from the item or the performance.

In this analysis, Cronbach's "alpha coefficient" may be defined as the ratio of the variability stemming from performances to the total variability. The higher this ratio is, the more we can be assured that the ratings reflect differences among the 3,924 instructional performances rather than some other sources. We may think of this ratio as a "signal-to-signal-plus-noise" ratio. The "signal" is the information we receive regarding performances--how high or low they are. The "signal-plus-noise" is this information as we discover it embedded within these data, i.e. masked to some degree by irrelevant or unknown sources of variation. The question reduces to this: How clear and unambiguous is the signal?

When this analysis was carried out, the alpha coefficient was found to be .95, which is extremely high. At the same time, it was found that deleting any one of the 13 items from the total scale reduced alpha by less than .01. Thus, we may conclude that Part I of the USES is internally consistent. To the extent that Part I treated as a single scale is similar to Factor 1, we may conclude that individual scores on Factor 1 having a high degree of reliability (i.e. internal consistency) are obtainable from these data.

At this writing there is no comparable evidence for Factor 2. Since Factor 2 appears to be less critical than Factor 1, rather than delay this report further, I decided to skip over this question with only a few comments. The main point is that Factor 2 is bound to have a much lower reliability for at least three interrelated reasons. First, it is much less clear what the second factor is and how to interpret it. Second, relatively few items are found in Part 1 which make up the second factor. Third, the correlations among the items which make up the second factor tend to be quite modest. For all of these reasons, it is unlikely that we would find much internal consistency in any subscale designed to measure the second factor as it is presented earlier in this report.

Stability of USES (Part A)

There is another meaning to reliability besides internal consistency which is often called stability. This is the extent to which the survey will yield basically the results obtained for all instructors who were rated at Time 1 at other times as well. For the same reasons as given above, the question will be examined only in reference to Factor 1. For the sake of convenience, data for item 13 (perceived overall effectiveness) alone will be examined. Although item 13 is arguably the best single indicator of Factor 1, there is certainly some loss in substituting one item for all of the items that correlate with the factor. However, this loss should tend toward a conservative estimate of stability, and it is perhaps best to err in a conservative direction.

Toward this end, 90 instructors were identified for whom annual summaries of USES scores were available for 1981-82, 1982-83, 1983-84, and 1984-85. These data made possible three pairwise comparisons of 1981-82 scores on item 13 with later score distributions (N = 90). Therefore three separate Pearson correlations were performed. For 1981-82 versus 1982-83, the two distributions of scores correlated .84. The other two correlations were .77 and .59 for comparisons separated by two and three years, respectively.

These results reveal that there is at least a modest amount of stability, but this decreases dramatically over a three year period. On reflection, had we found an extremely high level of stability, this would be a quite discouraging prospect for an instructor seeking to increase the level of his or her overall performance (not merely on item 13, but on all items which correlated with Factor 1). This would mean that nothing short of a miracle could effect such a change.

Having said this much, there is absolutely no evidence that the average score for these 90 instructors either improved or worsened over the 3-year period. The means varied less than three hundredths of a point! However, these modest coefficients of stability do suggest that there is hope for improvement, particularly among that crucial minority of new instructors who had not yet reached a plateau of performance when these data were obtained.

One of the factors which may have limited the level of stability on the USES is the attrition of many instructors who have performed at an unacceptably low level. This undoubtedly contributed to non-renewal of contracts for regular faculty as well as failure to re-hire wage faculty. The upshot of this attrition is a "restriction of range" effect. When the range of levels or scores on some ability or test score is limited through attrition, selection or some other factor, what occurs is an artificial reduction of correlations between that measure of ability and any other correlated ability or score. In effect, one is looking at only a small segment in a bivariate distribution because of some selection bias in the data itself. This rationalization of findings is important to bear in mind not only when considering stability, but particularly when estimating validity.

HOW VALID IS USES (PART A)?

The final major question is how valid is USES (Part 1) i.e. (to what extent) what is being measured is suitable for the purposes for which it is used.

If the major portion of variation over some 50,000 ratings of nearly 4,000 courses is demonstrably related to "instructor's perceived classroom effectiveness," we may reasonably conclude that the survey produces data that are pertinent to a major objective. In a great deal of the correspondence in which data from USES are referred to, it is implicitly assumed that perceived instructor effectiveness is the major issue. Few question that the data are pertinent to this objective. We have already produced evidence to support this widely held assumption.

However, this evidence is not yet adequate to answer the question of how valid the survey is. There needs to be some analytical basis for determining whether the data are "valid enough" for the various uses to which they actually are, or might be, put. For example, are summary scores for an individual suf-

ficiently valid to provide one basis for a decision to recommend or not recommend for promotion, tenure, special recognition or whatever. In the following, I will try to provide some evidence that bears upon this issue together with an analytical approach toward this class of questions.

In one sense, student evaluations of instructor performance have an immediate, or face validity. These data are extremely important in their own right in the same sense that customer self-reports on their satisfaction with a product should be taken seriously and given the highest priority. Thus, it makes little sense to dismiss the data as "mere opinion" since it is precisely these "opinions" or perceptions which, compounded over time, can have a lot to do with the reputation of a department or of the college as a whole. Similarly, specific items in USES (Part 1) can provide definite suggestions for needed changes in instruction. While other evidence may be helpful, these data speak for themselves and can have an application.

Nevertheless, I believe more than face validity is necessary if the USES data are to be used confidently and wisely. We need to know as much as possible about how USES results compare with professional peer assessment of teaching performance, or any other reflections upon instruction which are independent of (i.e. not influenced by) USES results. Independent evidence can make one more informed about the advantages and possible disadvantages of relying on USES. To the extent all criteria are imperfectly related to possible success in teaching (which is inevitable), one must take into account the risks of two possible kinds of errors: 1) an unfavorable action derived from an estimate of performance which is too low (false alarm), and 2) a favorable action based on an estimate of performance which is too high (miss). Misses and false alarms are unavoidable simply because all measures are fallible. On the other hand, decisions must be made in spite of this fallibility. By using independent evidence we can estimate the degree of invalidity in any one measure. This has direct bearing on how much of a problem these errors will be.

However, other factors besides the degree of validity have an effect on these error rates (denoted by "false alarm" and "miss"). These errors are also influenced by preselection and by more or less arbitrary decisions as to how the scores will be used. The problem of preselection arises here because instructors who are evaluated on USES make up a non-random sample in that they are recruited and selected on the basis of criteria, some of which pertain to teaching effectiveness. The more effective this process, the better the average performance of instructors should be. Furthermore, instructors' evaluated at any point in time are mostly made up of "survivors," poorer teachers having been screened out. This very selectivity will create a problem, referred to earlier as "restriction of range." In essence, the lower ranges of teaching effectiveness have been eliminated from consideration. This will necessarily lower the correlation between the predictor (USES ratings) and the criterion because the potential bivariate distribution of scores has been truncated.

The problem of decision rules and how they affect the two types of errors noted above calls attention to the fact that any decision involves the risk of error, and minimizing the risk of one type of error often increases the risk of another type.

Figure 2 illustrates both of these basic problems by presenting USES data paired with an abstract criterion, "success in teaching." The problem of restriction of range is shown by distinguishing between all possible instructors

from an applicant pool (dots) and those persons who are actually evaluated in a given year (circles). This illustration depicts a quite restricted range. Whatever the multiple criteria by which the candidates were preselected, to the extent teaching effectiveness is related to these criteria, a valid measure of perceived instructional effectiveness will reduce the apparent relationship between evaluations (e.g., USES) and success in teaching. When we consider the distribution of dots, there is clearly a linear relationship such that the higher the evaluations, the higher the probability of success in teaching. But when the circles alone are considered a very different situation results in which there is no apparent relationship between evaluations and teaching success. This is probably an exaggeration of the situation that actually exists. But it is quite certain that there is some restriction of range, and therefore any validity estimates found will tend to be artifically low.

Figure 2: Hypothetical data illustrating the effect of preselection and cut-offs on decision error rates

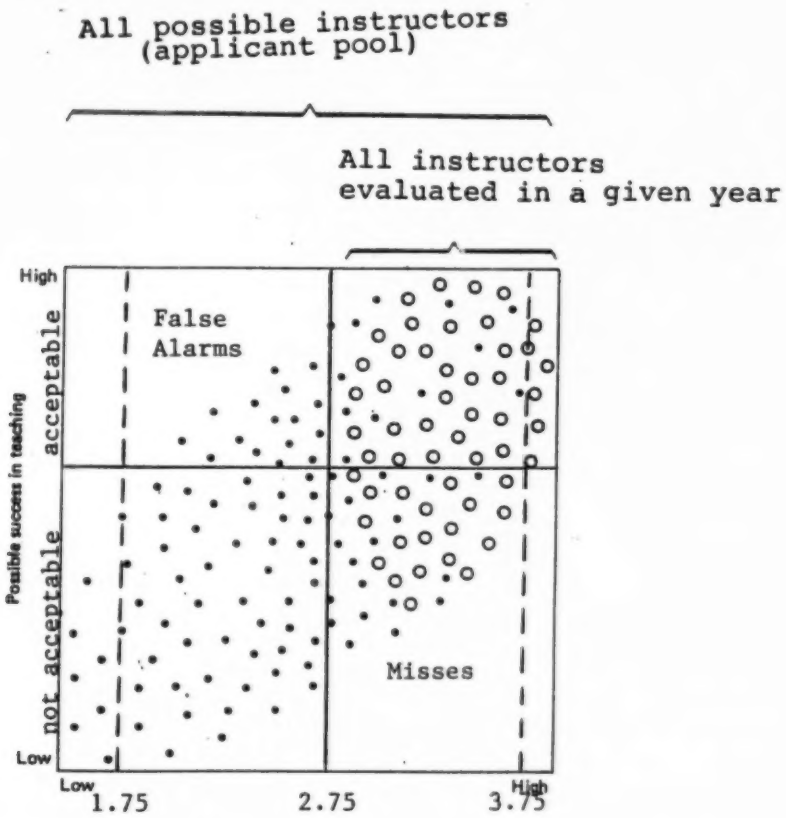


Figure 2 assumes that teaching success is measurable in some way, and the horizontal line represents the decision as to whether the teacher is acceptable or not. The question that now arises is what happens to the occurrence of errors when a valid evaluation instrument is used as the basis for predicting ac-

ceptability and different cut-off scores on this instrument are used. Since decisions are dichotomous in nature (accept/reject, pass/fail) whereas Figure 2 shows a continuous bivariate distribution, errors are a necessary consequence of arbitrary cut-offs or decision points. However, the main point to be illustrated here is that raising or lowering a cut-off score reduces one type of error at the expense of another.

Compare three possible cut-off points on the evaluation instrument in terms of what happens to the occurrence of false alarms (upper left-hand quadrant) and misses (lower right-hand quadrant). These are 1.75, 2.75 and 3.75. With a very high cut-off of 3.75, there are no misses whatsoever, but a large percentage are misclassified as probable failures in teaching when they actually would succeed. This appears to be too stringent a criterion. When the cut-off is very low (1.75), the reverse happens: there are no false alarms, but many misses. That is, many are misclassified as probable successes who turn out to be failures. This criterion is too lax. There are logically no other possibilities except something in between (say, 2.75), where a balance is struck between the two types of errors. At 2.75, one has chosen to live with a certain amount of each kind of error. However, there is no clearcut reason why 2.75 rather than some other intermediate number should be chosen as the cut-off. It is a matter of preference as to which error type should be most controlled.

Figure 2 should not be construed as depicting the actual validity of the USES. However, certain data will be presented which have a direct bearing on validity, but which fall short of yielding an overall picture of USES validity. By combining these results with the abstract view which Figure 2 was intended to provide, a reasonable basis for conclusions on validity, it is hoped, will be made possible.

There is no simple, readily available measure of the abstract "success in teaching" which forms the ordinate of Figure 2. In particular, it is impossible to correlate scores on an instructor evaluation instrument with success for all levels of success, due to the problem of preselection. A more accessible approach is to examine extreme cases of success or its lack, from distinguished teaching on the one hand to teachers with problems on the other. If these extreme cases can be identified, independently of USES, and shown to have either high USES scores (if distinguished) or low scores (if persons with problems), this approach would tend to support the validity of the survey.

"Validity" is used somewhat loosely here, since the survey was never intended, nor is it ever used, to identify distinguished teachers. However, it is reasonable to examine evaluation data for instructors independently identified as outstanding. The argument would be that if there is in any sense a continuum of possible success in teaching, an instrument which can predict high distinction should also serve well in predicting notable lack of success. This argument is crucial given the fact that the opportunity to test the ability to predict poor performance is limited due to the selected nature of the sample of evaluees. It is also quite reasonable to look for outstanding instructors since they should be relatively easy to identify in much the same way that 7-footers are easy to spot in a tribe full of 5-footers.

Following this approach, a search was conducted to discover as many as possible of the instructors on the regular faculty who have been presented with awards either partially or wholly based on distinguished performance as an instructor. A small number of such awards have been made. It was assumed that

student evaluations were not a major source of information on which the decisions to give the awards were based. In some cases, the awardees were recognized before the USES began to be administered. It is also reasonable to assume that outstanding instructors tend to stay outstanding from year to year. Thus, regardless of the year of awards annual average USES scores were found for all awardees for the four years beginning in fall, 1981. Of course, entries are missing for years in which the awardee was not employed here since evaluation data were not available. One awardee left before USES began and is therefore missing altogether.

These data are presented in Table 10.

Table 10. Annual average USES scores for eleven professors* recognized by awards for their outstanding achievements.

	Evaluation Years			
Professor	1981-82	1982-83	1983-84	1984-85
A	3.42	3.45	3.44	3.45
B	3.43	3.43	3.41	3.50
C	3.39	3.54	3.42	3.40
D	3.47	3.49	3.39	3.50
E	3.69	3.80	3.76	3.63
F	3.41	3.28	3.42	3.29
G	3.69	3.64	3.18	3.40
H	3.52	3.44	--	--
I	3.68	3.76	3.75	3.75
J	3.69	3.60	3.37	3.40
K	3.56	3.45	3.62	3.61
All college Ave-1)**	3.17	3.13	3.12	3.15
All college Ave-2)***	3.23	3.20	3.23	3.24

*The eleven professors received thirteen awards.
**The first all college average includes all instructors who were evaluated during these years, adjuncts as well as regular faculty.
***The second all college average includes only 80 instructors on the regular faculty for whom data were found for all four years.

These data reveal that these distinguished professors had extremely high USES scores. Their averages, year by year, were: 3.54, 3.53, 3.48 and 3.49. These averages are about one standard deviation above the all college average for 80 regular faculty who were evaluated between 1981 and 1985. (Although statistical tests of significance could have been done, these differences were so substantial that a test of significance is superfluous.) This comparison is with "survivors" who are, by and large, probably successful teachers. This record of sustained outstanding scores on USES by those who have been recognized by students and professional peers as outstanding supports the validity of overall USES evaluations. Apparently these scores do tap some of the same dimensions that go into recognition for high quality contributions to teaching.

To further substantiate the point that high USES scores are related to professional judgments of peers, the following series of quotes was taken from the

peer evaluations of nine of the eleven honored professors. (These quotes give no clues as to their exact identities.)

- "a superior teacher and a very effective teacher"
- "extremely effective in teaching these [difficult to teach] courses"
- "demonstrates abundant enthusiasm for teaching"
- "popular and very enthusiastic classroom teacher"
- "one of the most effective persons on the CNC faculty"
- "teaching effectiveness continues to be excellent"
- "impressive command of his subject matter"
- "attained excellence in every dimension"
- "truly impressive record as a teacher"

Although it is not difficult to locate positive comments among peer evaluations, it is clear in these instances that the peer members have reached for superlatives to describe these outstanding professors: "extremely effective," "superior," "truly impressive," etc.

One caveat is that these individuals are certainly not the only professors at CNC with impressive records as teachers. However, the selection seems likely to represent among the most "outstanding" teachers, where that adjective cannot be applied to very many more professors without losing all meaning. All may be good, but all are not "outstanding."

Although high evaluations appear to be related to independent professional assessments of performance, it is more difficult to predict who will be a poor performer. The main problem, of course, is that poor performance in teaching frequently leads to termination. The difficulty is compounded when one tries to guarantee that the identification of instructors with problems is not influenced by evaluation results. With the USES data, this requires looking back at evidence before 1981, when USES was adopted. After fall, 1981, any judgments of problem performance could have been influenced by actual USES results. One must avoid circularity in predicting low USES scores from a basis of low USES scores; this circularity does not support validity. Pre-1981 evidence of low teaching evidence therefore seems required from a logical standpoint. However, this restriction has its own problems. One is simply that those whose instructional performance before 1981 was marginal tend to have departed from among us. Secondly, there is no available data that consistently measures the success of teaching apart from student evaluations. Other evidence is haphazard, anecdotal and inconsistent.

This leaves the simple test of survival. If an instructor has left, one of the reasons for this could be substandard performance of duties as an instructor. However, there also exist a host of other reasons why an instructor might leave: to accept another position at higher pay, to move to Arizona for a spouse's health, etc., etc. Once all of these other reasons are ruled out, we

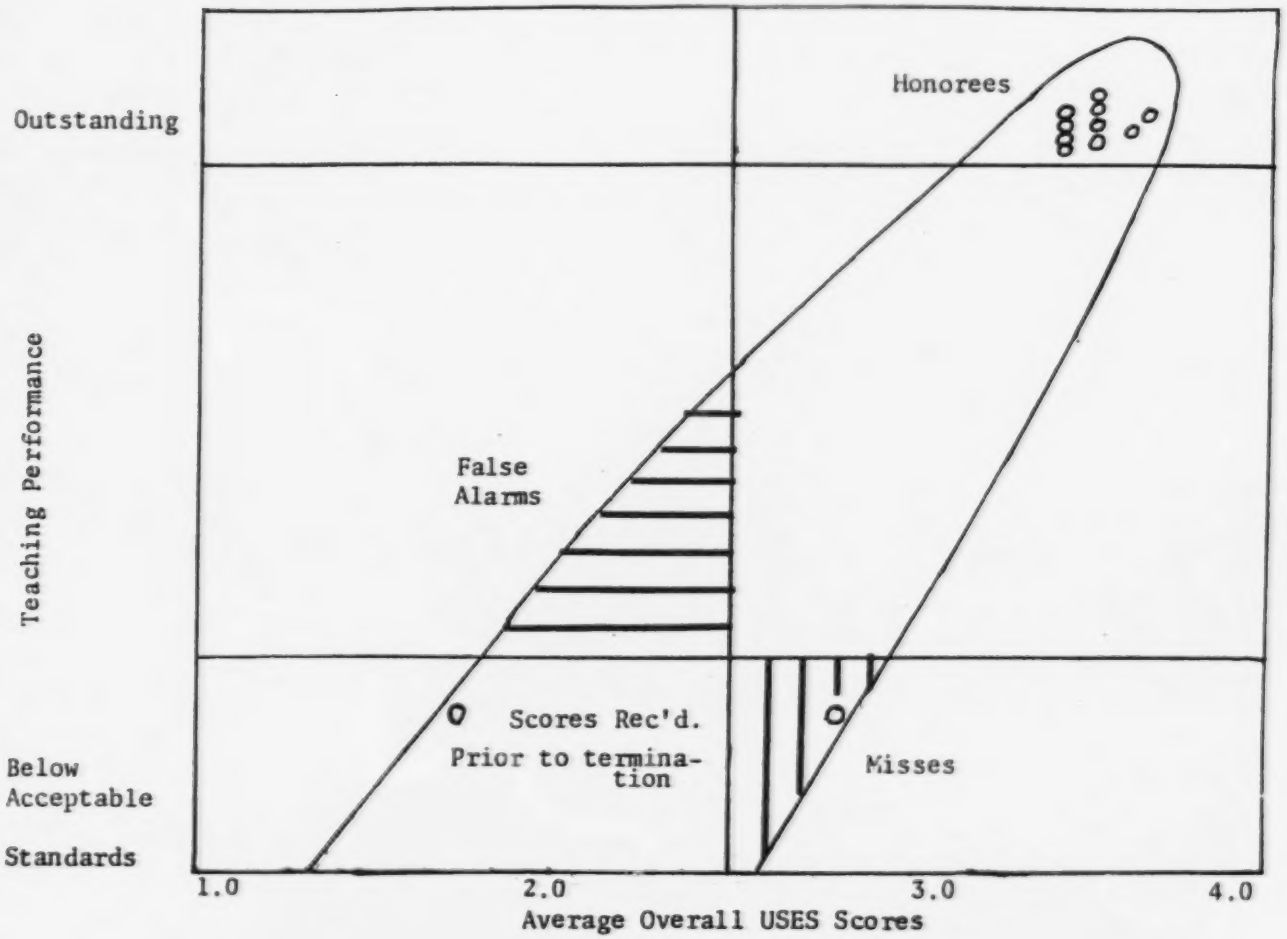
should have a small residue of instructors whose non-survival is attributable to poor instructional performance. If we choose these from among those who were on the staff in 1981 and had one or two rounds of USES evaluations, although the decision not to renew might have been influenced to a small degree by the evaluations, in all likelihood the decision was based on much more than this evidence.. This was the rationale for comparing faculty members who were on staff in fall, 1981 but who are no longer on staff with those who have been continuously on staff since 1981, where poor performance as an instructor was the main reason given by the department chairman for non-continuance. These comparisons were in terms of average overall performance on USES evaluations.

On the regular faculty, only one individual could be identified who met the criteria. The average annual overall evaluations of this individual, from fall, 1981 to termination, were 2.76 and 1.74. These averages are well below the all-college annual averages which were around 3.2 for those years.

For adjunct faculty, a similar comparison could not be made. Although there were several instructors employed in 1981 who are no longer employed, and the reasons specifically relate to performance, in a number of cases the reason given was directly tied to low evaluations by students. This casts doubt on such evidence since it is circular. There appears to be no escape from this circularity since adjunct professors are not evaluated by the same kind of detailed process which applies to regular faculty. Therefore decisions are likely to be influenced largely by evaluation data. In contrast, the regular faculty member with low evaluations also presented other evidence that contributed to the decision.

A conservative summary of the available data on validity would be that they reveal the expected relationship between USES scores and more or less independent assessments of perceived excellence as a classroom instructor. However, the evidence is meager. As far as it goes, it is consistent with the following hypothetical figure (Figure 3), which assumes that: (a) those who are independently identified as "outstanding" have very high USES scores, and (b) as USES scores fall below about 2.5 the chances of termination, on the grounds of inadequate performance as a teacher, increase dramatically. The figure also reveals an important consequence which flows from these assumptions: the risks of error in basing decisions (promotion, tenure, etc.) on USES alone are quite substantial. Both major kinds of errors would be found frequently, i.e. misses and a false alarms. The latter include professors with low USES scores who still "make the grade" as teachers. The former are those whose USES scores are not terribly low but who are still found to be unacceptable.

Figure 3. Hypothetical Relationship Between Overall USES Scores and Teaching Performance ("False alarms" and "Misses" defined by an arbitrary cut-off score of 2.5.)



This summary should merely reinforce the trend of permitting USES only as supplemental input for making these key decisions; they should under no circumstances be the sole arbiter. The risks of such unadvised use are particularly great for individuals. Whereas even a modest degree of validity can be useful from the institutional perspective, since the gains will offset the errors in the long run, these "errors" are more than mere imperfections in the system for the individual. An "error" for the individual can mean a serious career setback. For the institution the loss of one potentially fine instructor is not a tragedy. For the individual it can be.

SUMMARY OF MAJOR CONCLUSIONS

1. This report had a limited objective. It did not address concerns related to the administration of USES, the timing and frequency of evaluations, the diagnostic use of the results, demographic items, etc. The sole objective was to address concerns as to reliability, validity, and what the survey consistently measures.

2. Principal components analysis of previous USES results reveals that "perceived classroom effectiveness" is consistently what is measured by the survey. Item 13 "overall instructor performance," is the best single item measure of this factor.

3. The other items which share the most variance with overall classroom effectiveness are quite similar from one department to another.

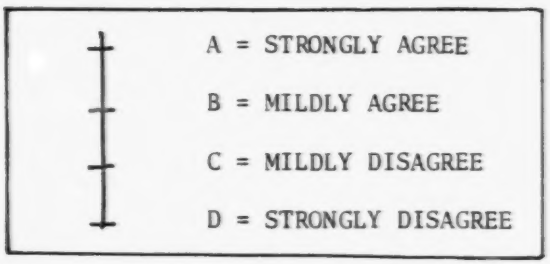
4. USES results show a high degree of internal consistency. That is, separate parts of the survey are well correlated with the survey as a whole (i.e. Part A).

5. USES results also suggest that scores are fairly stable over time, with stability dropping sharply after about two years.

6. USES results are generally related to more or less independent assessments of excellence as a classroom teacher. However, the probable risks of error which would arise from USES scores as a sole basis for important decision are significant and substantial. This observation should reinforce using the scores cautiously to supplement other critical evidence.

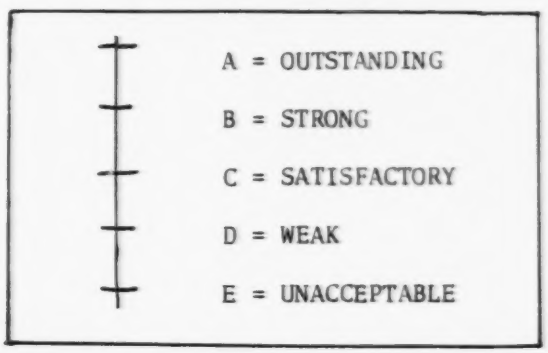
APPENDIX: Uniform Student Evaluation Survey (Part A)

Directions: Please record your responses for this section in the appropriately numbered blanks on your response sheet. In reacting to each of the first twelve statements, please use the following scale for your responses.



If you have no opinion regarding some statement or if you believe that it does not apply to this class, please leave its response space blank.

1. The course the instructor presented was consistent with the course description in the Catalogue.
2. The instructor came to class well-prepared.
3. The instructor's grading policies were clearly explained early in the term.
4. The instructor's grading procedures were fair.
5. Graded assignments and tests were returned promptly.
6. Tests covered knowledge, application, or reasoning that could be expected on the basis of course lectures and readings.
7. The instructor maintained regular office hours and/or encouraged students to seek help when needed.
8. The instructor demonstrated command of the subject matter of the course.
9. The instructor's presentation of the course material was clear and understandable.
10. The instructor demonstrated interest in and enthusiasm for the subject matter of the course.
11. The instructor was intellectually motivating and stimulated learning.
12. The instructor met classes on time and for adequate duration.
13. At the end of many courses, the overall performance of students is often graded on a scale of the following kind:



Which of these ratings do you consider to be the most accurate description of the overall instructional performance of the instructor in this course?